Hybrid acoustic models for distant and multichannel large vocabulary speech recognition

Swietojanski, P., Ghoshal, A., & Renals, S. (2013). HYBRID ACOUSTIC MODELS FOR DISTANT AND MULTICHANNEL LARGE VOCABULARY SPEECH RECOGNITION. In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). IEEE.

HYBRID ACOUSTIC MODELS FOR DISTANT AND MULTICHANNEL LARGE VOCABULARY SPEECH RECOGNITION

Pawel Swietojanski, Arnab Ghoshal and Steve Renals

Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB

{p.swietojanski,a.ghoshal,s.renals}@ed.ac.uk

ABSTRACT

We investigate the application of deep neural network (DNN)-hidden Markov model (HMM) hybrid acoustic models for far-field speech recognition of meetings recorded using microphone arrays. We show that the hybrid models achieve significantly better accuracy than conventional systems based on Gaussian mixture models (GMMs). We observe up to 8% absolute word error rate (WER) reduction from a discriminatively trained GMM baseline when using a single distant microphone, and between 4–6% absolute WER reduction when using beamforming on various combinations of array channels. By training the networks on audio from multiple channels, we find the networks can recover significant part of accuracy difference between the single distant microphone and beamformed configurations. Finally, we show that the accuracy of a network recognising speech from a single distant microphone can approach that of a multi-microphone setup by training with data from other microphones.

Index Terms— Distant Speech Recognition, Deep Neural Networks, Microphone Arrays, Beamforming, Meeting recognition

1. INTRODUCTION

Distant Speech Recognition (DSR) [1] remains a significant open challenge. Recognition of speech captured using multiple distant microphones, typically configured in a calibrated array, is a difficult task since the speech signals to be recognised are degraded by the presence of other acoustic sources and by the effects of reverberation. However, there has been progress in distant speech recognition over the past decade, with a particular focus on the transcription of multiparty meetings captured using tabletop or wall-mounted microphones, stimulated by the NIST Rich Transcription (RT) evaluation campaigns [2].

Most distant speech recognition systems have adopted a two-part architecture in which a microphone array beamforming algorithm is applied to the recorded multichannel speech, followed by conventional acoustic modelling approaches. Good examples of such systems include the AMIDA [3] and ICSI/SRI [4] systems for meeting transcription. Both of these systems process the microphone array signals using a noise-reducing Wiener filter on each channel, followed by delay-sum beamforming where the time delays of arrival are estimated using generalized cross-correlation with phase transform (GCC-PHAT) [5] and smoothed using a two-stage Viterbi post-processing [6]. The beamformed audio may then be processed in the same way as single channel speech, typically using speech activity detection (if the recording is not already segmented), followed by a speech recogniser.

More sophisticated beamforming algorithms have been proposed that take into account the correlation of the noise on different channels under spherically isotropic or cylindrically isotropic noise field assumptions. Such approaches, collectively referred to as superdirective beamforming [7], work well for speech enhancement by improving directional selectivity at lower frequencies. However, such techniques are designed for generic sounds and as such they neither take into account the unique characteristics of human speech, nor are designed specifically to improve speech recognition performance. There has been some work on designing a beamformer specifically assuming that its output will be used for speech recognition. For instance: the maximum negentropy beamformer [8] exploits the fact that the distribution of the subband samples of clean speech is super-Gaussian whereas the distribution of noise-corrupted speech is closer to Gaussian; LIMABEAM (likelihood maximising beamforming) [9] optimises the array processing parameters to maximise the likelihood of the recognised hypothesis given the filtered acoustic data. LIMABEAM may be thought of as explicitly optimising the beamforming to maximise speech recognition accuracy by taking acoustic model likelihood as a surrogate for accuracy.

Extracting a single enhanced channel does not address the problem of overlapped speech. Hori et al. [10] describe a system which applies a dereverberation algorithm to the multichannel audio, followed by a source separation approach comprising a speaker diarisation component based on clustered direction-of-arrival estimates, which is then used to direct a delay-sum beamformer.

This research was supported by EPSRC Programme Grant grant, no. EP/I031022/1 (Natural Speech Technology). Thanks to Erich Zwyssig for providing beamforming scripts; Jonathan Kilgour for help with AMI corpus annotations; and Thomas Hain for helpful discussions.

Several researchers have explored ways to perform recognition from multiple distant microphones without performing explicit beamforming. Wölfel et al. [11] have investigated approaches in which each individual channel is separately recognised, with the recognition hypotheses combined using confusion network combination. A variant of this approach also recognises an enhanced channel obtained by beamforming, which is then added to the confusion network combination. Stolcke [12] has investigated this approach in detail on a meeting recognition task, concluding that combining the individual channels at the signal level by delay-sum beamforming is superior (in terms of both accuracy and processing time) compared to the individual channel approach. Marino and Hain [13] have performed some initial investigations training GMM-based systems on concatenated feature vectors from 2–4 microphones. This produced encouraging word error rates, similar to those obtained by beamforming the signals from the same microphones.

In this paper we investigate the use of deep neural networks (DNN) for distant speech recognition. We do this in the context of a meeting recognition task using the AMI corpus[1] [14], which is a collection of meetings recorded at three sites in Europe (Edinburgh, UK; IDIAP, Switzerland; TNO, Netherlands). In this paper we explore two novel alternatives to conventional beamforming for speech recognition using multiple distant microphones:

1. Simple concatenation — the DNN performs feature-level combination using a single input feature vector from the concatenation of the individual feature vectors from each microphone channel channel;
2. Multi-style training — the DNN is trained and tested using a single distant microphone, but is trained on the outputs of multiple array channels, taken one at a time.

We compare these approaches to conventional beamforming, in which the DNN is trained on a single beamformed channel, analogous to the systems discussed above [3, 4], and to systems that use a single distant microphone. We also compare the DNN systems to a discriminatively trained conventional GMM-based system.

The motivation for applying DNNs to recognition of meetings recorded with distant microphones is twofold. Firstly, DNNs are a powerful framework for learning representations [15] from multiple sources of information, acting as a cascade of nonlinear feature extractors (followed by a log-linear classifier). This offers the possibility of a less constrained feature-level combination compared with beamforming. Secondly, DNNs have been shown to not only provide significantly improved recognition accuracies over GMM-based recognisers [16], they have also been found to subsume the effects of different compensation schemes commonly used in GMM-based systems [17]. Thus, using DNNs for meeting recognition we envision to have a simple yet accurate single system instead of the multi-pass decoding and re-estimation that is often applied for distant speech recognition [3, 8, 18].

[1] http://corpus.amiproject.org/

2. DNN-HMM HYBRIDS FOR SPEECH RECOGNITION

In a deep neural network-hidden Markov model hybrid system [19, 20, 16], the DNN is trained to classify the input acoustics into classes corresponding to the HMM states. After training, the output of the DNN is an estimate of the posterior probability $P(s|\mathbf{o}_t)$ of each state s given the acoustic observations $\mathbf{o}_t$ at time t. The computation performed by the network may be written as:

$$\begin{aligned}\mathbf{u}_l &= \sigma(\mathbf{W}_l\mathbf{u}_{l-1} + \mathbf{b}_l), \quad \text{for } 1 \le l < L \\ \mathbf{a}_L &= \mathbf{W}_L\mathbf{u}_{L-1} + \mathbf{b}_L, \\ P(s|\mathbf{o}_t) &= \frac{\exp\{a_L(s)\}}{\sum_{s'}\exp\{a_L(s')\}},\end{aligned}$$

where $\mathbf{u}_l$ is the input to the $l+1$-th layer, with $\mathbf{u}_0 = \mathbf{o}_t$; $\mathbf{W}_l$ is the matrix of connection weights between $l-1$-th and l-th layers; $\mathbf{b}_l$ is the additive bias vector at the l-th layer; $\mathbf{a}_L$ is the activation at the output layer; and $\sigma(x) = 1/(1+\exp(-x))$ is a sigmoid non-linearity, also known as the activation function. The recogniser uses a pseudo log-likelihood of state s given observation $\mathbf{o}_t$:

$$\log p(\mathbf{o}_t|s) = \log P(s|\mathbf{o}_t) - \log P(s),$$

where $P(s)$ is the prior probability of state s calculated from the training data [19].

We use stochastic gradient descent (SGD) to train DNNs, minimising a negative log posterior probability cost function over the set of training examples $\mathcal{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$:

$$\theta^* = \arg\min_\theta -\sum_{t=1}^{T}\log P(s_t|\mathbf{o}_t),$$

where $\theta = \{\mathbf{W}_1, \ldots, \mathbf{W}_L, \mathbf{b}_1, \ldots, \mathbf{b}_L\}$ is the set of parameters of the network, and s_t is the most likely state at time t obtained by a forced-alignment of the acoustics with the transcript. This is also the expected cross-entropy between the distribution represented by the reference labels and the predicted distribution. We use an unsupervised pretraining phase using greedy layer-wise training of RBMs [21] to initialise the DNN parameters before training them using SGD.

3. EXPERIMENTAL SETUP

For our experiments, we use the AMI corpus[1], which contains around 100 hours of meetings recorded in specifically equipped instrumented meeting rooms at three sites in Europe (Edinburgh, IDIAP, TNO). There are two types of meetings—scenario based, where four speakers act out certain predetermined roles of a design team (project manager, designer,

etc.), as well as non-scenario-based which are natural spontaneous meetings on a range of topics. The scenario-based meetings make up about 70% of the corpus. Each meeting usually has four participants and the meetings are in English, albeit with a large proportion of non-native speakers. Acoustic signal is captured by multiple microphones including individual head microphones (IHM), lapel microphones, and one or more microphone arrays. Each recording site uses a primary 8-microphone uniform circular array of 10 cm radius, as well as a secondary array whose geometry varies between sites. In this work we use the primary array and refer to it as the multiple distant microphones (MDM) variant. Experiments with single distant microphone (SDM) make use of first microphone of the primary array.

Most previous research using the AMI corpus [3, 22] have done so in the context of the NIST RT evaluations, where the AMI data was used together with other meeting corpora. In order to perform more controlled experiments with identical microphone array configurations, we have defined a 3-way partition of the AMI corpus into train, development, and test sets[2]. This partition makes about 78 hours of speech available for training, and holds out about 9 hours each for development and test sets. All the three sets contain a mix of scenario- and non-scenario-based meetings, and are designed such that no speaker appears in more than one set. The definitions of these sets have also been made available on the AMI corpus website[1]. We use the segmentation provided with the AMI corpus annotations (version 1.6). In this work, we consider all segments (including those with overlapping speech), and the speech recognition outputs are scored by the `asclite` tool [23] following the NIST RT[3] recommendations for scoring simultaneous speech.

3.1. Acoustic models

For the IHM configuration, 7 frames (3 on each side of the current frame) of 13-dimensional MFCCs (C0-C12) are spliced together and projected down to 40 dimensions using linear discriminant analysis (LDA) and decorrelated using a single semi-tied covariance (STC) transform [24]. These features are referred to as LDA+STC. Both the GMM-HMM and DNN-HMM acoustic models are speaker adaptively trained (SAT) on these LDA+STC features using a single feature-space maximum likelihood linear regression (FMLLR) transform estimated per speaker. The GMM-HMM systems provide the state alignments for training the DNNs. Additionally, the DNNs are trained on 40-dimensional log Mel filterbank (FBANK) features appended with delta and acceleration coefficients. The state alignments used for training the DNNs on FBANK features are the same as those used for the LDA+STC features. Through some initial pilot experiments we found the DNNs trained on LDA+STC transformed MFCCs to produce around 1.5-2% (absolute) lower word error rates (WERs) when compared to those trained on MFCCs and roughly the same results as those trained on FBANK features.

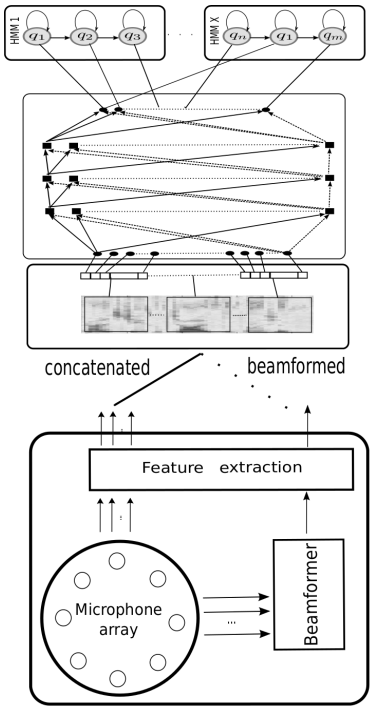

Fig. 1. Front-end for our setups with DNN in hybrid configuration on the top.

For the audio captured using the distant microphones, a Wiener filter-based noise cancellation is first applied using the Qualcomm-ICSI-OGI front-end tools [25]. For the MDM experiments, we follow the noise cancellation with a delay-sum beamforming on either 2, 4, or 8 uniformly-spaced array channels using the BeamformIt toolkit [6]. In both the SDM and MDM case, the audio (noise-cancelled and beamformed, respectively) is then processed in a similar fashion to the IHM configuration. The major difference between the IHM and SDM/MDM configurations is that when audio is captured with distant microphones, it is not realistically possible to ascribe a speech segment to a particular speaker without using speaker diarisation. As such, the SDM/MDM experiments do not use any form of speaker adaptation or adaptive training. In our pilot experiments we did not see a consistent advantage from adapting to the entire meeting.

The GMM-HMM systems are trained on the speaker adapted LDA+STC features for the IHM case, or on the unadapted features for the SDM/MDM case, using the boosted maximum mutual information (BMMI) [26] criterion. The number of tied-states are roughly 4000 in all configurations,

[2] `http://www.cstr.inf.ed.ac.uk/reproducibleResearch/Swietojanski-ASRU-2013/index.html`

[3] `http://nist.gov/speech/tests/rt/2009`

Table 1. Word error rates (%) for the GMM and DNN acoustic models for various microphone configurations.

System	Microphone configurations				
	IHM	MDM8	MDM4	MDM2	SDM
Development set					
GMM BMMI on LDA+STC	29.4 (SAT)	54.8	56.5	58.0	63.2
DNN on LDA+STC	26.7 (SAT)	51.4	51.5	51.6	55.4
DNN on FBANK	28.3	51.1	-	52.9	55.8
Evaluation set					
GMM BMMI on LDA+STC	31.6 (SAT)	59.4	61.2	62.9	67.6
DNN on LDA+STC	28.4 (SAT)	56.0	55.9	56.5	59.8
DNN on FBANK	31.5	55.6	-	57.9	60.8

and each of the GMM-HMM systems have a total of 80,000 Gaussians. These are then used to provide the state alignments for training the corresponding DNNs using either the LDA+STC features or the FBANK features. The GMM-HMM systems are trained using the Kaldi speech recognition toolkit [27], while the DNNs are trained using in-house tools based on the Theano library [28] and running on general-purpose graphics processing units.

Following our previous experience and those reported by others [17, 29, 16, 30, 31], the DNNs were configured to have 6 hidden layers with 2048 neurons in each hidden layer. The network parameters are initialised from stacked restricted Boltzmann machines (RBMs) that are pretrained in a greedy layer-wise fashion [21]. The networks are trained using SGD following an exponentially decaying learning schedule. The various hyper-parameters are same as those described in [30], except the initial learning rate, which was tuned to 0.06.

3.2. Lexicon and language model

We use the same 50,000 word AMI pronunciation dictionary used in [3]. An in-domain trigram language model (LM) is estimated from the 801K words of the training transcripts, which is then interpolated with two other trigram LMs—one estimated from 3M words of the Switchboard training transcripts, and the other from 22M words of the Fisher English transcripts. The LMs are estimated using interpolated Kneser-Ney smoothing. The in-domain AMI LM has an interpolation weight of 0.73, the Fisher LM gets a weight of 0.22, while the contribution from the Switchboard LM is negligible with a weight of 0.05. The final interpolated LM has 1.6M trigrams and 1.5M bigrams, and achieves a perplexity of 78 on the development set.

4. RESULTS

As described in Section 3.1, we use three MDM configurations where beamforming is done on 2, 4, and 8 channels respectively. The results obtained by both the BMMI-trained GMM system and the DNN systems for these configurations, as well as for the SDM and IHM conditions are shown in Table 1. The WERs for the GMM-based systems are comparable to the ones reported previously in [13, 3] on AMI-based test sets, albeit using different training-test partitions.

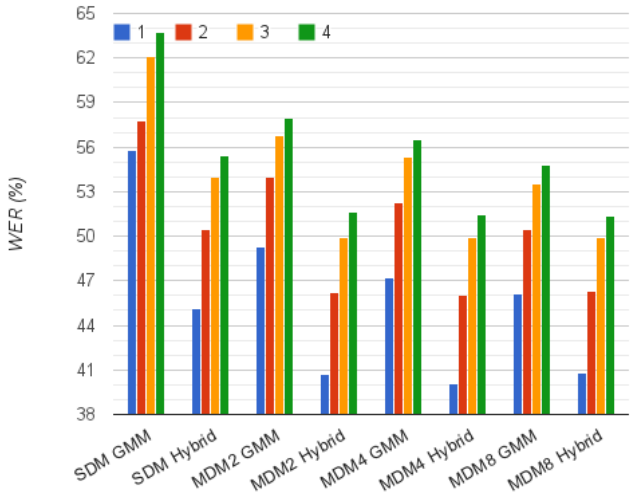

Fig. 2. Development set WERs for segments with 1, 2, 3 and 4 overlapping speakers. AMs are trained on MFCC LDA+STC features.

We find the DNNs to greatly improve recognition accuracy for speech recorded with distant microphones. In fact, the network trained on SDM data is only 0.6% absolute worse than the best GMM-BMMI system built from beamformed audio from 8 far-field microphones. Interestingly enough, the DNNs are also found to be less sensitive to the number of beamformed channels used, particularly with the LDA+STC features. We attribute this to the fact that multiple layers of non-linear transformations can better compensate against small variabilities in feature space [32].

While Table 1 presents the WER for all segments, including those with overlapped speech, Figure 2 shows the WERs for segments with different numbers of overlapped speakers. As one may expect, overlapped segments are harder to recognise. In fact, even if a beamformer can select the dominant source perfectly it still does not address the problem of recognising overlapped speech. Figure 2 gives us a sense of the

Table 2. WER on channels 1 & 2 for the DNNs trained on multiple channels (MFCC LDA/STC features). SDM models are trained on channel 1.

Combining method	Recog. Channel	Devset	Evalset
SDM (no combination)	1	55.4	59.8
SDM (no combination)	2	55.1	59.8
Concatenate 1+5	1	52.9	57.5
Concatenate 1+3+5+7	1	52.7	57.3
Multi-style 1+3+5+7	1	52.8	57.5
Multi-style 1+3+5+7	2	52.7	57.7

difficulty in recognising overlapped speech. We see a 8-12% reduction in WER when only considering segments with non-overlapping speech.

Through a second set of experiments, we evaluate the extent to which a DNN is able to learn to do the front-end processing—both noise-cancellation and beamforming—by providing the features extracted from multiple microphones as input to the networks. In these initial experiments the networks still have 6 hidden layers like in the previous case[4] except with a wider input layer. Note that this is not entirely comparable to the setup where the DNNs are trained on features extracted from beamformed audio, since the Wiener filtering and beamforming are time domain operations, whereas the DNNs trained on concatenated features are operating entirely in cepstral or log-spectral domains. Nevertheless, the results give us an indication of how complementary the information in different channels are. We see from Table 2 and Figure 3 that the DNNs trained on concatenated inputs do in fact work substantially better than the SDM case, and achieve results approaching that of the beamformed configurations. The important point to note here is that the DNNs trained on concatenated features do not use any knowledge of the array geometry. Consequently, the technique, just like the approach of [13], is applicable to any arbitrary configuration of microphones.

To further understand the nature of the compensation being learned by the DNNs with multi-channel inputs, we do an additional control experiment. The input to the DNN is from a single channel, and at test time this is identical to the SDM case. However, during training the data from other channels are also presented to the network, although not at the same time. In other words, the DNN is presented with data from channel 1, channel 3, and so on, in successive mini-batches during training, while at test time it is only tested on a single channel. We call this the multi-style training, and it is related to our previous work [33], where the same basic concept was used to train DNNs in a multilingual fashion. From Table 2 we see that this approach performs similarly to the DNNs with concatenated input. Recognition results on channel 2, which is not used in the multi-style training, show similar trends.

[4]However, since the networks are being tasked with additional processing, deeper architectures may be more suitable.

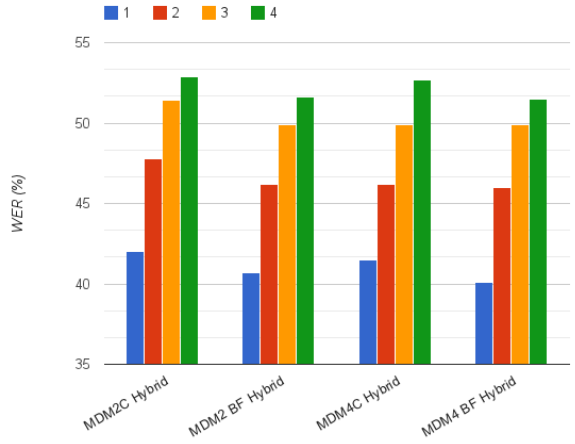

Fig. 3. Comparison of DNNs trained on concatenated (C) channels to the ones trained on noise-cancelled and beamformed (BF) signals for segments with 1, 2, 3, and 4 overlapping speakers.

These results strongly suggest that there is information in a single channel to have more accurate recognition. However, extraneous factors in the data may confound a learner trained only on data from a single channel. Being forced to classify data from multiple channels using the same shared representation (i.e. the hidden layers) the network, almost by definition, has to ignore the channel-specific covariates. To the best of our knowledge, this is the first result to show that it is possible to improve recognition of audio captured with a single distant microphone by guiding the training using data from microphones at other spatial locations.

5. CONCLUSION

In this work we presented some promising results on using hybrid DNN-HMM models for distant speech recognition with a single distant microphone or multiple distant microphones. We show that it is possible to improve recognition results by concatenating the features extracted from multiple channels and proving that as input to a DNN. This is applicable in cases where the array geometry is unknown. More interestingly, we show that it is possible to improve the recognition with a single distant microphone to the same level as that of multiple distant microphones by only constraining the training with multi-microphone data.

6. REFERENCES

[1] M Wölfel and J McDonough, *Distant Speech Recognition*, Wiley, 2009.

[2] J Fiscus, J Ajot, and J Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *Multimodal Technologies for Perception of Humans*, R Stiefelhagen, R Bowers, and J Fiscus, Eds., number 4625 in Lecture Notes in Computer Science Volume, pp. 373–389. 2008.

[3] T Hain, L Burget, J Dines, PN Garner, F Grezl, AE Hannani, M Huijbregts, M Karafiat, M Lincoln, and V Wan, "Transcribing meetings with the AMIDA systems," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 486–498, 2012.

[4] A Stolcke, X Anguera, K Boakye, O Cetin, A Janin, M Magimai-Doss, C Wooters, and J Zheng, "The SRI-ICSI Spring 2007 meeting and lecture recognition system," in *Multimodal Technologies for Perception of Humans*, R Stiefelhagen, R Bowers, and J Fiscus, Eds., number 4625 in Lecture Notes in Computer Science Volume, pp. 373–389. 2008.

[5] CH Knapp and GC Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320–327, 1976.

[6] X Anguera, C Wooters, and J Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 7, pp. 2011–2021, 2007.

[7] J Bitzer and KU Simmer, "Superdirective microphone arrays," in *Microphone Arrays*, M Brandstein and D Ward, Eds., pp. 19–38. Springer, 2001.

[8] K Kumatani, J McDonough, B Rauch, D Klakow, PN Garner, and W Li, "Beamforming with a maximum negentropy criterion," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 5, pp. 994–1008, 2009.

[9] M Seltzer and R Stern, "Subband likelihood-maximizing beamforming for speech recognition in reverberant environments," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, pp. 2109–2121, 2006.

[10] T Hori, S Araki, T Yoshioka, M Fujimoto, S Watanabe, T Oba, A Ogawa, K Otsuka, D Mikami, K Kinoshita, T Nakatani, A Nakamura, and J Yamoto, "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 2, pp. 499–513, 2012.

[11] M Wölfel, C Fügen, S Ikbal, and J McDonough, "Multi-source far-distance microphone selection and combination for automatic transcription of lectures," in *Proc ICSLP*, 2006.

[12] A Stolcke, "Making the most from multiple microphones in meeting recognition," in *Proc IEEE ICASSP*, 2011.

[13] D Marino and T Hain, "An analysis of automatic speech recognition with multiple microphones," in *INTERSPEECH*, 2011, pp. 1281–1284.

[14] S Renals, H Bourlard, J Carleta, and A Popescu-Belis, *Multimodal Signal Processing*, Cambridge University Press, 2012.

[15] Y Bengio, A Courville, and P Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[16] G Hinton, L Deng, D Yu, GE Dahl, A-R Mohamed, N Jaitly, A Senior, V Vanhoucke, P Nguyen, TN Sainath, and B Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[17] F Seide, G Li, X Chen, and D Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. IEEE ASRU*, 2011.

[18] K Kumatani, J McDonough, and B Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 127–140, 2012.

[19] H Bourlard and N Morgan, *Connectionist Speech Recognition—A Hybrid Approach*, Kluwer Academic, 1994.

[20] S Renals, N Morgan, H Bourlard, M Cohen, and H Franco, "Connectionist probability estimators in HMM speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 161–174, 1994.

[21] G Hinton, S Osindero, and Y Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[22] F Grézl, M Karafiát, S Kontár, and J Černocký, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Proc. ICASSP*, 2007, vol. 4, pp. IV–757–IV–760.

[23] JG Fiscus, J Ajot, N Radde, and C Laprun, "Multiple dimension Levenshtein edit distance calculations for evaluating automatic speech recognition systems during simultaneous speech," in *Proc. LREC*, 2006.

[24] MJF Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Trans. Speech, Audio Process.*, vol. 7, no. 3, pp. 272–281, 1999.

[25] A Adami, L Burget, S Dupontb, H Garudadric, F Grezl, H Hermansky, P Jain, S Kajarekar, N Morgan, and S Sivadas, "Qualcomm-ICSI-OGI features for ASR," in *In Proc. ICSLP*, 2002, pp. 21–24.

[26] D Povey, D Kanevsky, B Kingsbury, B Ramabhadran, G Saon, and K Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. IEEE ICASSP*, 2008, pp. 4057–4060.

[27] D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlíček, Y Qian, P Schwarz, J Silovský, G Stemmer, and K Veselý, "The Kaldi speech recognition toolkit," in *Proc. IEEE ASRU*, 2011.

[28] J Bergstra, O Breuleux, F Bastien, P Lamblin, R Pascanu, G Desjardins, J Turian, D Warde-Farley, and Y Bengio, "Theano: a CPU and GPU math expression compiler," in *Proc. SciPy*, 2010.

[29] GE Dahl, D Yu, L Deng, and A Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, 2012.

[30] P Swietojanski, A Ghoshal, and S Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. IEEE Workshop on Spoken Language Technology*, Miami, Florida, USA, Dec. 2012.

[31] K Veselý, A Ghoshal, L Burget, and D Povey, "Sequence-discriminative training of deep neural networks," in *Proc. INTERSPEECH*, 2013.

[32] M Seltzer, D Yu, and Y Wang, "An investigation of deep neural networks for noise robust speech recognition," in *In Proc. ICASSP*, 2013.

[33] A Ghoshal, P Swietojanski, and S Renals, "Multilingual training of deep neural networks," in *Proc. IEEE ICASSP*, 2013, pp. 7319–7323.